Section 1:

My patent teaches a configuration of a "second set" of "volume tracking cameras" responsive to a "first set" of "area tracking cameras." Jain only teaches of a "set of volume tracking cameras."

References from my patent in support of this teaching:

From the Background of the Invention:

Top of page 6:

1- "The system employs a matrix of separate overhead tracking cameras responsible for first locating any given object as a whole in a local (X, Y) area rather than in a (X, Y, Z) volume coordinate system. This technique yields a substantially uniform pixel resolution per area tracked providing a simple and regular approach to camera arrangement when the system is scaled to track larger areas.

2- The system employs separate sets of one or more pan, tilt, and zoom cameras per player to be tracked. These moveable cameras are automatically directed by the system based upon the (X, Y) location information that was first determined using the overhead tracking "area" cameras. Each of these volume cameras will collect (X, Y, Z) information from a particular view of the player to be combined with at least the (X, Y) information captured by the "area" cameras concerning the same player. Due to the system's ability to move and zoom each player-tracking camera, a substantially uniform pixel resolution per player is achieved. This technique provides a simple and regular approach to camera arrangement when the system is scaled to track more and more players."

From the Summary of the Invention:

Bottom of page 14:

"For vision-based systems such as Motion Analysis and Vicon ***(and Jain – they all use "Volume Tracking Camera Arrangements")*** significant difficulties also begin to present themselves in consideration of the requirement that each joint be in view of at least two cameras at all times. As players move about and change their body positions, individual joints can easily be lost from view (inclusion*) or take positions that from a given camera's "flat 2D" perspective appear to make them a part of a different player. This is especially true in light of the small spherical markers used by existing systems that are not in view from every possible rotational angle of the joint. What is needed is a vision-based tracking system that can identify and track players with one set of cameras and then automatically direct a second set of cameras to adjust their views so as to minimize these inclusions*. What is further needed is a tracking system that employs markers that can cover a much larger area, for instance all the way around an elbow rather than a single point or set of three points upon the elbow, while at the same time remaining less obtrusive."

** Note: In my patent I incorrectly used the word "inclusion" when I meant "occlusion."(This mistake was made throughout the original text.)*

Middle of page 16:

"The present inventors prefer to separate the function of "area" tracking for 2D movements and identification from "player" tracking for the full 3D data set. Hence, a matrix of overhead (X, Y) cameras follows the movement of any number of players per single camera across a single playing area. Since the number of markers tracked from the overhead view is limited to primarily the helmet and shoulders, the total number of marks, even for a large number of players is still trackable. Furthermore, due to the overhead view and the tendency for players to remain upright, it is expected that there will be minimal instances of helmet or shoulder "overhead inclusions." In addition to being tracked in 2D space by the overhead cameras, each player will also be followed by at least two and preferably four dedicated pan, tilt, and zoom perspective cameras. These dedicated, movable, perspective cameras will be automatically directed by the known (X, Y) location of each player as determined by the overhead cameras. This combination of novel techniques provides for scalability by both tracking area and player. Hence, to cover more area simply add one or more overhead "area" tracking cameras while to cover more players simply add additional sets of one or more dedicated movable "player" tracking cameras."

Bottom of page 34 through top of page 35:

"As previously discussed, the present inventors prefer the novel approach of separating the (X, Y) tracking of each player or game object as a whole from the (X, Y, Z) determination of each player's critical locations (e.g., joints and body parts). Furthermore, the novel concept of locating all of the identity markings on the "upper surface" of each player facilitates the top-oriented field of view of the (X, Y) tracking assemblies which naturally limits inclusions due to player bunching. Given this separation of tasks, the present invention becomes highly scalable. Each overhead (X, Y) area tracking assembly covers a fixed and uniform square or rectangular tracking area. Furthermore, the pixel resolution per this area remains substantially constant. To cover more tracking area, simply add additional (X, Y) assemblies. Each player that moves throughout this connected tracking area has their helmets (located at a minimum) and ideally also their shoulders. From this information, the player is identified along with pertinent information including orientation, direction of movement, velocity, and acceleration as well as current relative (X, Y) location. Using the current (X, Y) location as well as the direction of movement, velocity and acceleration, the preferred embodiment controllably directs one or more (X, Y, Z) pan, tilt and zoom cameras to automatically follow the given player. Furthermore, by intelligent inspection of the various projected player paths, the system optionally switches (X, Y, Z) cameras from one player to another to best maximize overall tracking performance. By constantly zooming each (X, Y, Z) camera to maximize player size per field of view, a uniform and ideal pixel resolution per marker square inch is maintained. To cover more players, simply add additional (X, Y, Z) pan, tilt and zoom camera sets per added player."

From the Objects and Advantages:

Bottom of page 38:

"Further objects and advantages are to provide:

1- a system that is scalable and therefore comprises uniform assemblies that are combinable into a matrix designed to increase tracking coverage in terms of area, volume and the number of objects while still maintaining uniform performance;"

From the Description of the Drawings:

Top of page 41:

Figs. 13a and **13b** depict the theory and implementation of fixed (X, Y, Z) volume tracking camera assemblies. ***(Jain's approach)***

Figs. 14a and **14b** depict the theory and implementation of fixed (X, Y) area tracking camera assemblies. ***(My "first set of area tracking cameras")***

Figs. 15a and **15b** depict the theory and implementation of movable (X, Y, Z) volume tracking camera assemblies. ***(My "second set of movable volume tracking cameras", responsive to the "first set")***

From the Detailed Description:

Middle of page 45 through top of page 47:

"Referring now to **Fig. 13a**, there is shown an example of a fixed (X, Y, Z) volume tracking camera **502** that comprises a camera **126**, filter and connection to a local computer system for video processing and analysis **160**. Camera **126** can be one of any analog or digital-imaging cameras as typically used for industrial vision applications. One example is the Eagle digital camera used by Motion Analysis Corporation that features a ceramic metal oxide semiconductor (CMOS) image sensor with 1280 x 1024 pixel resolution and a maximum capture rate of 600 million pixels per second. It is important to note that, once in place, this volume tracking camera **126** has a fixed field-of-view (FOV) similar to a four-sided pyramid in shape within an image cone **121v**. The actual pixel resolution per inch of the FOV will vary throughout the height **121h** of the pyramid ranging from a higher value at the top width **121tw** to a lower value at the bottom width **121bw**. These cameras are typically secured from an overhead position to have a perspective view arrangement **502m** of the desired tracking volume as shown in **Fig. 13b**.

Referring now to **Fig. 13b**, there is shown one particular arrangement of fixed (X, Y, Z) cameras **502** that, when taken together, form a uniquely shaped tracking volume through which a player **17**, wearing markers such as spherical markers **17sm**, may transverse. ***(This is Jain's approach.)*** The resultant resolution per cross-sectional area of this volume **121tv** is non-uniform. For example, while skating through any given point in the tracking volume, markers **17sm** on one body part of player **17** may be viewed by camera **126e** with a much lower resolution per inch than similar markers on a different body part. Also, the second camera such as **126d** may have a much different pixel resolution of marker **17sm** than camera **126e**. Cameras **126a**, **126b** and **126c** may each have obstructed views of marker **17sm**.

Referring now to **Fig. 14a**, there is shown an example of fixed (X, Y) area tracking camera **504**, that comprises a tracking camera **124** with a filter **124f** that have been enclosed in a protective housing **121** with a transparent underside **121a**. Also enclosed in housing **121** is an energy source **10** emitting tracking energy **11** as well as unfiltered filming camera **125**. Tracking camera **124** and filming camera **125** are connected to a local computer system for video processing and analysis **160**. The entire assembly included within housing **121** is preferably secured in an overhead position looking directly down at a subset of the tracking surface. From this overhead position, camera **124** has a fixed FOV

120v that is focused on the top surface of any players below and as such maintains a substantially uniform pixel resolution per tracking area FOV **120v**.

Referring now to **Fig. 14b**, there is shown a scalable area tracking matrix **504m** comprising multiple fixed (X, Y) area tracking cameras **120c** aligned such that their FOVs **120v** are substantially side-by-side with a small overlap for calibration purposes. Throughout this scalable matrix **504m**, the top surface **110** of player **17** can be readily tracked.

Referring now to **Fig. 15a**, there is shown an example of movable (X, Y, Z) volume tracking camera **506**, that comprises a pan, tilt and zoom camera **140** with a filter that is connected to local computer system for video processing and analysis **160**. Top surface **110** of player **17** is held in constant view by one or more of cameras **140** that are controllably panned, tilted and zoomed for maximum desirable pixel resolution per player. The information for this controlled movement is based either upon the current (X, Y) coordinates of player **17** as previously determined from information gather by scalable area tracking matrix **504m** or by movement tracking algorithms calculated by computer **160** to predict the next possible location of player 17.

Referring now to **Fig. 15b**, there is shown a scalable volume tracking matrix **506m** comprising multiple movable volume tracking cameras **506** where one or more cameras form an assembly and are dynamically assigned to a player 17. As will be explained in more detail using **Figs. 16a** and **16b**, this dynamic process of automatically panning, tilting and zooming each movable camera to maintain the maximum desirable pixel resolution per player provides a substantial benefit over the arrangement of fixed volume tracking cameras **502**."

Middle of page 41:

"**Fig. 16b** depicts how player bunching and therefore marker inclusion* is addressed by the use of movable (X, Y, Z) volume tracking camera assemblies."

Note: This word should be "occlusion," not "inclusion."

Middle of page 47:

"Referring now to **Fig. 16b**, there is shown an example matrix of four FOV's **120v** created by area tracking cameras **124**. Within this combined grid, several players having top surfaces such as **110x** and **111x** move freely about. In this particular example, four movable cameras **140-a**, **140-b**, **140-c** and **140-d** are tracking the player with top surface **110x**. As depicted, the FOV's for cameras **140-b** and **140-d** are almost fully blocked by other players whereas the FOV for camera **140-a** is partially blocked but the FOV for camera **104-c** is clear. The preferred embodiment will automatically reassign cameras such as **141-d** that may already be tracking another player, (e.g., the player with top surface **111x**) to now follow a different player with top surface **110x** so as to ensure total maximum player visibility. This reassignment decision can be based upon the information gathered by the scalable area

tracking matrix **504m**, predictive calculations made by computer **160** concerning the expected next positions of any and all players, or both."

Middle of page 60:

"Referring now to **Fig. 3**, there is shown a block diagram depicting all of the major components **1000** of which a subset has been identified as the preferred embodiment **1004** of the present invention. Preferred embodiment **1004** comprises fixed (X, Y) area tracking assemblies **504** in combination with movable (X, Y, Z) volume tracking assemblies **506**,..."

Top of page 61:

1- "By limiting the fixed (X, Y) area tracking matrix to a top view only, the system creates a scalable approach to camera placement that provides a substantially uniform pixel resolution per area;

2- By implementing a separate matrix of movable (X, Y, Z) volume tracking cameras to pick up the remaining side views of the players, the system creates a scalable approach to camera placement that provides a substantially uniform pixel resolution per player;"

The following references are taken from Jain's patent and show his "set of volume tracking cameras" as distinctly different from my "second set" of "volume tracking cameras" responsive to a "first set" of "area tracking cameras."

Abstract: (consecutive lines)

1. "Multiple video cameras, each at a different spatial location, produce multiple two-dimensional video images of the real-world scene, each at a different spatial perspective."
2. "Objects of interest in the scene are identified and classified by computer in these two-dimensional images."
3. "The two-dimensional images of the scene, and accompanying information, are then combined in the computer into a three-dimensional video database, or model, of the scene."

Col 5, Lines 46-51:

1. "The MPI video system of the present invention will be seen to avail itself of multiple two-dimensional video images from each of multiple stationary cameras as are assembled into a three-dimensional video image database (an important element of the present invention)."

Col 8, Lines 49-60:

2. "The method of the invention is directed to presenting to a user/viewer a particular, viewer-selected, two-dimensional video image of a real-world, three-dimensional scene.
3. In order to do so, multiple video cameras, each at a different spatial location, produce multiple two-dimensional images of the real-world scene, each at a different special perspective.
4. Objects of interest in the scene are identified and classified in the two-dimensional images.
5. These multiple two-dimensional images of the scene, and their accompanying object information, are then combined in a computer into a three-dimensional video database, or model, of the scene."

Col 10, Lines 17-21:

1. "Forth, the user/viewer-specified criterion may be of a particular object in the scene. In this case the computer will combine the images from the multiple video cameras ... to generate a three-dimensional video model of the scene..."

Col 11, Line 66 – Col 12, Line 7:

2. "Fig. 13 is a graphical illustration showing the intersection formed by the rectangular viewing frustum of each camera scene onto the environment volume in the GM-PPS portion of the MPI video system of the present invention; the filled frustum representing possible areas where the object can be located in the 3-D model while, by use of multiple views, the intersection of the frustum from each camera will closely approximate the 3-D location and form of the object in the environment model."

Col 14, Lines 54-59:

3. "It is likely a better idea to construct a comprehensive video image database from quality images obtained from only a few strategically positioned cameras, and to then permit universal construction of customized views from this database, all as is taught by the present invention."

Col 17, Lines 3-8:

4. "The high level architecture for a MPI video system so functioning is shown in a first level block diagram in Fig. 1. An image at a certain perspective from each camera 10a, 10b, ... 10n is converted to its associated camera scene in camera scene buffers CSB 11a, 11b, ... 11n. Multiple camera scenes are then assimilated into the environment model 13 by computer process in the Environ. Model Builder 12."

Col 22, Line 63 – Col 23, Line 7:

1. "Three dimensional Scene Analysis. The purpose of scene analysis is to extract three-dimensional information from video frames captured by cameras. This process is performed in the following two stages:"
2. "First, 2-D information is extracted. From each video frame, feature points such as players and field marks are extracted and a list of feature points is generated."
3. "Second, 3-D information is extracted ***(from the same original data used to extract the 2-D information – this comment added by me)***. From the two-dimensional description of the video frame, three-dimensional information in the scene, such as player position and camera status, is then extracted."

Col 27, Lines 37-41:

1. "In the present approach an omniscient multi-perspective perception system uses multiple stationary cameras which provide comprehensive coverage of an extended environment. The use of fixed global cameras simplifies visual processing."

Col 32, Lines 43-47:

2. "One of the goals of the exercise of the multi-perspective perception system was to illustrate the advantages of using static cameras for scene capture, and the relative simplicity of visual processing in the scenario when compared to processing from a single camera."

Col 37, Lines 59-64:

1. "It is the teaching of the present invention how to so construct from multiple two-dimensional video images a three-dimensional database, and how to so manage the three-dimensional database for the production of two-dimensional video images that not necessarily those images from which the database was constructed."

In summary, I believe that it is clear that all of the 2-D tracking information and 3-D video display information created by Jain's proposed system is coming from a single set of "multiple video cameras, each at a different spatial location." Furthermore, all of these cameras are fixed.

My patent clearly teaches the first creation of the 2-D tracking information using a first set of aligned area tracking cameras followed by the second creation of the 3-D display information using a second set of controllably movable volume tracking cameras, responsive to the 2-D tracking information.

With respect to the critical nature of this difference between my patent and Jain, I provide the following explanation taken from may patent (middle of Page 30 through middle of Page 31.)

> "The major reason for their difficulties" ***(i.e. with the Jain approach)*** "is the "volume tracking" approach they have taken to solving this problem. Simply put, each contestant may move about the playing surface which because of his or her own height and the expected flight of any game objects, becomes a playing volume rather than an area. Every portion of this volume must be in view of two or more cameras at all times. The cameras must be fixed prior to the contest so that they may be calibrated as a network. If players will have a tendency to bunch up, additional cameras will be needed within any given volume to create additional views thereby reducing anticipated inclusions. The cameras must be limited in the field of view so that they can maintain a sufficient resolution or pixels per inch within their field in order to detect the reflected markers. As the playing area widens, it becomes increasingly difficult to place cameras close enough to the inner volumes so that the ideal field of view is maintained per camera without causing an obstruction to the players or viewing audience. This obstruction would occur if a mounting structure were created to hang the cameras directly above the inner volumes. Newer cameras will continue to provide higher resolutions theoretically allowing the cameras to move further back from any given volume and still maintain the requisite pixels per inch resolution. As cameras pull back, however, the distance between the energy tracking source, the reflective marker and the cameras will continue to increase thereby having a negative effect on signal strength.
>
> The present inventors prefer to separate player (object) tracking into two distinct sub-processes thereby eliminating the aforementioned problems. In distinct contrast to the "volume tracking" approach, the preferred embodiment of the present invention relies upon a "player following" controlled by an "area tracking" technique. In essence, the players are first tracked in two dimensions, X and Y, throughout the playing area. The currently determined location of each player is then used to automatically direct that player's individual set of cameras that "follow" him or her about the playing surface. This two-step approach has many critical advantages when faced with tracking objects throughout a larger volume. First, locating the "top surface" (i.e., helmet) of each player in X-Y space for substantially the entire contest is significantly simpler than trying to detect their entire form from two or more cameras throughout the entire playing volume. Second, by placing the player id on their "top surface" (i.e., their helmet and if need be shoulder pads), the system is able to easily identify each player while it also tracks their X-Y coordinates. Third, by controllably directing one or more automatic pan, tilt and zoom cameras to follow each player (and game object) the ideal field of view and maximum resolution can be dynamically maintained per player."

I respectfully submit that this is a substantial difference in our teachings and based upon this alone, I argue that the Jain patent should not be considered prior art.

Section 2:

My patent teaches that each camera in the "first set" of "area tracking cameras" should be aligned in a regular matrix with every other camera in that set, such that their individual fields-of-view are:

a. substantially perpendicular to the tracking area;

b. substantially parallel with each other, and

c. sufficiently overlapping to create a single contiguous, substantially coplanar view of the tracking area.

References from my patent in support of this teaching:

From the Background of the Invention:

Top of page 5:

"The present inventors have addressed many of these drawbacks in their co-pending applications entitled:

- Multiple Object Tracking System, Application Serial No. 09/197,219; Filed: November 20, 1998
- Method for Representing Real-Time Motion over the Internet, Application Serial No. 09/510,922; Filed: February 22, 2000
- Employing Electromagnetic By-Product Radiation for Object Tracking, Application Serial No. 09/881,430; Filed: June 14, 2001
- Visibly Transparent Wide Observation Angle Retroreflective Materials, Application Serial No. 09/911,043; Filed: July 23, 2001

Each of these patent applications is hereby incorporated by reference into the present application.

In these patents applications, the present inventors describe various aspects of a multiple object tracking system that functions in general to track many types of objects but that is especially constructed to track athletes during a live sporting event such as an ice hockey game. These patent applications teach at least the following novel components:

1. The system employs a matrix of separate overhead tracking cameras responsible for first locating any given object as a whole in a local (X, Y) area rather than in a (X, Y, Z) volume coordinate system. This technique yields a substantially uniform pixel resolution per area tracked providing a simple and regular approach to camera arrangement when the system is scaled to track larger areas."

(Note: The "Multiple Object Tracking System" patent contains significant discussion and figures relating to the preferred configuration and alignment of the first set of overhead area tracking cameras.)

Bottom of page 34 through top of 35:

"As previously discussed, the present inventors prefer the novel approach of separating the (X, Y) tracking of each player or game object as a whole from the (X, Y, Z) determination of each player's critical locations (e.g., joints and body parts). Furthermore, the novel concept of locating all of the identity markings on the "upper surface" of each player facilitates the top-oriented field of view of the (X, Y) tracking assemblies which naturally limits inclusions due to player bunching. Given this separation of tasks, the present invention becomes highly scalable. Each overhead (X, Y) area tracking assembly covers a fixed and uniform square or rectangular tracking area. Furthermore, the pixel resolution per this area remains substantially constant. To cover more tracking area, simply add additional (X, Y) assemblies."

Bottom of page 46:

"Referring now to **Fig. 14a**, there is shown an example of fixed (X, Y) area tracking camera **504**, that comprises a tracking camera **124** with a filter **124f** that have been enclosed in a protective housing **121** with a transparent underside **121a**. Also enclosed in housing **121** is an energy source **10** emitting tracking energy **11** as well as unfiltered filming camera **125**. Tracking camera **124** and filming camera **125** are connected to a local computer system for video processing and analysis **160**. The entire assembly included within housing **121** is preferably secured in an overhead position looking directly down at a subset of the tracking surface. From this overhead position, camera **124** has a fixed FOV **120v** that is focused on the top surface of any players below and as such maintains a substantially uniform pixel resolution per tracking area FOV **120v**.

Referring now to **Fig. 14b**, there is shown a scalable area tracking matrix **504m** comprising multiple fixed (X, Y) area tracking cameras **120c** aligned such that their FOVs **120v** are substantially side-by-side with a small overlap for calibration purposes. Throughout this scalable matrix **504m**, the top surface **110** of player **17** can be readily tracked."

Top of page 53:

"Furthermore, it is expected that additional volume cameras **140** assigned to track the same player **17** will similarly be simultaneously calibrated with camera **124**. It should be noted that player **17** may be straddling a boundary between area tracking cameras **124** and as such two different volume cameras **140** may actually be calibrated for the same player **17** by two different area cameras **124**. In practice, this is immaterial since the pre-calibration by system **160** of the entire scalable area tracking matrix **504m** can be thought of as creating one large single area (X, Y) tracking camera. Hence, it can be seen that each of the volume cameras such as **140** in the present figure or **140-a**, **140-b**, **140-c** and **140-d** of prior figures that are currently assigned to follow player **17** are simultaneously calibrated frame-by-frame to the overhead matrix **540m**. Furthermore, once calibrated the multiple cameras such as **140-a**, **140-b**, **140-c** and **140-d** may be used to stereoscopically locate markings on player **17** that are not in view of the overhead matrix **540m**."

The following references are taken from Jain's patent and show that he teaches no particular alignment or arrangement for his single set of "multiple video cameras, each at a different spatial location."

Figures: Depicted Camera Arrangement

- Always showing:
    - Three to four cameras;
    - Each camera is fixed;
    - Each camera is at a different physical location *and orientation* with respect to each neighboring camera;
        - Hence, the cameras are not arranged in a regular configuration with their fields-of-view aligned in space to form a contiguous, coplanar field-of-view;
- See Fig.'s 2, 5, 8, 11a, 14a, 14b, 18 and 21

Abstract: (consecutive lines)

4. "Multiple video cameras, each at a different spatial location, produce multiple two-dimensional video images of the real-world scene, each at a different spatial perspective."

Col 5, Lines 46-51:

6. "The MPI video system of the present invention will be seen to avail itself of multiple two-dimensional video images from each of multiple stationary cameras as are assembled into a three-dimensional video image database (an important element of the present invention)."

Col 8, Lines 49-60:

7. "The method of the invention is directed to presenting to a user/viewer a particular, viewer-selected, two-dimensional video image of a real-world, three-dimensional scene.
8. In order to do so, multiple video cameras, each at a different spatial location, produce multiple two-dimensional images of the real-world scene, each at a different special perspective.

Col 14, Lines 54-59:

5. "It is likely a better idea to construct a comprehensive video image database from quality images obtained from only a few strategically positioned cameras, and to then permit universal construction of customized views from this database, all as is taught by the present invention."

Col 17, Lines 3-8:

6. "In such cases, the MPI video system of the present invention will, by use of the environment model and video synthesis techniques, synthesize a virtual camera, and video image, so as to view a scene, and episode, or an entire presentation from a view-specified perspective."
7. "The high level architecture for a MPI video system so functioning is shown in a first level block diagram in Fig. 1. An image at a certain perspective from each camera 10a, 10b, ... 10n is converted to its associated camera scene in camera scene buffers CSB 11a, 11b, ... 11n. Multiple camera scenes are then assimilated into the environment model 13 by computer process in the Environ. Model Builder 12."

Col 28, Lines 45-52:

3. "For the case where an object is observed by more than one camera, the three-dimensional voxel representation is particularly efficacious. Here a dynamic object recorded on an image plane projects into some set of voxels. Multiple views of an object will produce multiple projections, one for each camera,. The intersection of all such projections provides an estimate of the 3-dimensional form of the dynamic object as illustrated in Fig. 13 for an object seen by four cameras."

Col 34, Lines 43-50:

2. "As shown in Fig. 17, an object that is out of view, too small, and/or occluded from view in one camera is in view, large and/or un-occluded to the view of another camera. Note that the object labels used in the Fig. 17 are for explanation only. The prototype subsystem does not include any non-trivial object recognition, and all object identifiers that persist over time are automatically assigned by the system."

Col 35, Lines 24-29:

3. "The fifth view of Fig. 19e is a virtual view of the model from directly overhead the courtyard – where no real camera actually exists. The virtual view shows the exact locations of all three objects, including the robotic vehicle, in the two-dimensional plane of the courtyard."

In summary, I believe that it is clear from both the descriptions in the current patent, as well as the descriptions contained within the original continued application entitled "Multiple Object Tracking System," that I teach an arrangement of the first set of area tracking cameras that includes FOV's that are:

a. substantially perpendicular to the tracking area;

b. substantially parallel with each other, and

c. sufficiently overlapping to create a single contiguous, substantially planar view of the tracking area.

It is also clear that Jain only teaches, both in his figures and specification, that the multiple cameras should be at different locations and "spatial perspectives." These *different* spatial perspectives de facto eliminate camera FOV's that are:

a. substantially perpendicular to the tracking area;

b. substantially parallel with each other, and

since both of these two qualities require at least *substantially identical spatial perspectives*. Without such alignment, it is not possible to further combine FOV's such that they:

c. sufficiently overlapping to create a single contiguous, substantially planar view of the tracking area.

Section 3:

My patent teaches that the data from the "first" and "second" sets of cameras is combined to form a database comprising a multiplicity of 3-D coordinates for a multiplicity of key object features, such as body joints, etc., that requires minimal data storage and can be used to animate any desired 2-D viewing perspective. Jain teaches that the data from his "volume tracking cameras" is combined to form a database comprising 3-D video data that can be morphed into any desired 2-D viewing perspective.

References from my patent in support of this teaching:

From the Background of the Invention:

Top of page 5:

"The present inventors have addressed many of these drawbacks in their co-pending applications entitled:

- Multiple Object Tracking System, Application Serial No. 09/197,219; Filed: November 20, 1998
- Method for Representing Real-Time Motion over the Internet, Application Serial No. 09/510,922; Filed: February 22, 2000

- Employing Electromagnetic By-Product Radiation for Object Tracking, Application Serial No. 09/881,430; Filed: June 14, 2001
- Visibly Transparent Wide Observation Angle Retroreflective Materials, Application Serial No. 09/911,043; Filed: July 23, 2001

Each of these patent applications is hereby incorporated by reference into the present application.

(Note: The "Method for Representing Real-time Motion over the Internet" patent contains significant discussion and figures relating to the preferred 3-D kinetic body-point model collected by the system and used to animate a rendition of the tracking game.)

Top of page 13:

"A.1. Is the desired representation to be visual for display only or a mathematical model for measurement and rendering?

It is preferable that the tracking system creates a mathematical model of the tracked players and equipment as opposed to a visual representation that is essentially similar to a traditional filming and broadcast system. A mathematical model allows for the measurement of the athletic competition while also providing the basis for a graphical rendering of the captured movement for visual inspection from any desired viewpoint. Certain systems exist in the marketplace that attempt to film sporting contests for multiple viewpoints after which a computer system may be used to rotate through the various overlapping views giving a limited ability to see the contest from any perspective. All of the aforementioned machine vision companies, such as Motion Analysis and Vicon, generate a mathematical model."

Bottom of page 13:

"A.2. Is two or three-dimensional information preferred?
Three-dimensional information provides the ability to generate more realistic graphical renditions and to create more detailed statistics and analyses concerning game play. All of the aforementioned machine vision companies such as Motion Analysis and Vicon attempt to generate three-dimensional data while the beacon-based Trakus and Orad only generate two-dimensional information.

This requirement of creating a three-dimensional mathematical model of game play further dictates that at least the major joints on a player are identified and tracked. For instance, the player's helmet, shoulders, elbows, writs, torso, waist, knees, and feet are all beneficial tracking points for a 3D model. This informational goal in practice precludes beacon-based systems such as Trakus and Orad since it would require a significant number of beacons to be placed onto each player, often times in locations that are not convenient. Furthermore, each beacon will create additional signal processing and, given current state of the art in microwave tracking, the system could not function quickly enough to resolve all of the incoming signals.

For vision-based systems such as Motion Analysis and Vicon" ***(and Jain)*** ", significant difficulties also begin to present themselves in consideration of the requirement that each joint be in view of at least two cameras at all times. As players move about and change their body positions, individual joints can easily be lost from view (inclusion) or take positions that from a given camera's "flat 2D" perspective appear to make them a part of a different player. This is especially true in light of the small spherical markers used by existing systems that are not in view from every possible rotational angle of the joint. What is needed is a vision-based tracking system that can identify and track players with one set of cameras and then automatically direct a second set of cameras to adjust their views so as to minimize these inclusions. What is further needed is a tracking system that employs markers that can cover a much larger area, for instance all the way around an elbow rather than a single point or set of three points upon the elbow, while at the same time remaining less obtrusive."

Top of page 13:

"A.4. Must this information be collected and available in real time?

The ability to capture and analyze images, convert them into a 3-D mathematical model, and then dynamically render a graphic representation along with quantified statistics in real time offers significant opportunities and challenges. ..."

Top of page 33:

"As previously discussed, the present inventors prefer the novel approach of separating the (X, Y) tracking of each player or game object as a whole from the (X, Y, Z) determination of each player's critical locations (e.g., joints and body parts). Furthermore, the novel concept of locating all of the identity markings on the "upper surface" of each player facilitates the top-oriented field of view of the (X, Y) tracking assemblies which naturally limits inclusions due to player bunching. Given this separation of tasks, the present invention becomes highly scalable. Each overhead (X, Y) area tracking assembly covers a fixed and uniform square or rectangular tracking area. Furthermore, the pixel resolution per this area remains substantially constant. To cover more tracking area, simply add additional (X, Y) assemblies. Each player that moves throughout this connected tracking area has their helmets (located at a minimum) and ideally also their shoulders. From this information, the player is identified along with pertinent information including orientation, direction of movement, velocity, and acceleration as well as current relative (X, Y) location. Using the current (X, Y) location as well as the direction of movement, velocity and acceleration, the preferred embodiment controllably directs one or more (X, Y, Z) pan, tilt and zoom cameras to automatically follow the given player. Furthermore, by intelligent inspection of the various projected player paths, the system optionally switches (X, Y, Z) cameras from one player to another to best maximize overall tracking performance. By constantly zooming each (X, Y, Z) camera to maximize player size per field of view, a uniform and ideal pixel resolution per marker square inch is maintained. To cover more players, simply add additional (X, Y, Z) pan, tilt and zoom camera sets per added player."

Top of page 50:

"Referring now to **Fig. 22c**, there is shown information similar to **Figs. 21a**, **21b** and **21c** to dramatize **Fig. 22d** that depicts the image formed in computer **160** based upon non-uniform multi-shape **554** reflections. This collection of individual markings **17m** that have been placed at various locations on player **17** are only used to locate a particular body part and its orientation rather than to identify the player **17**. In the preferred embodiment that employs these types of flat **544** markings, the identification of player **17** is based upon a top surface of the body **564** Id Location **560**."

The following references are taken from Jain's patent and show that he teaches the use of 2-D video data that can be morphed into any desired 2-D viewing perspective" – as opposed to finding body-joints and creating a 3-D body-point model for later animation.

Abstract: (consecutive lines)

5. "Multiple video cameras, each at a different spatial location, produce multiple two-dimensional video images of the real-world scene, each at a different spatial perspective."
6. "Objects of interest in the scene are identified and classified by computer in these two-dimensional images."
7. "The two-dimensional images of the scene, and accompanying information, are then combined in the computer into a three-dimensional video database, or model, of the scene."

Col 5, Lines 46-51:

9. "The MPI video system of the present invention will be seen to avail itself of multiple two-dimensional video images from each of multiple stationary cameras as are assembled into a three-dimensional video image database (an important element of the present invention)."

Col 7, Lines 7-16:

10. "The present invention will be seen to manage a very great amount of video data. A three-dimensional video model, or database is constructed. For any sizable duration of video (and a sizable length thereof may perhaps not have to be retained at all, or at least retained long), this database is huge. More problematical, it takes very considerable computer "horse-power" to construct this database – however long its video data should be held and used."

Col 8, Lines 49-60:

11. "The method of the invention is directed to presenting to a user/viewer a particular, viewer-selected, two-dimensional video image of a real-world, three-dimensional scene.
12. In order to do so, multiple video cameras, each at a different spatial location, produce multiple two-dimensional images of the real-world scene, each at a different special perspective.
13. Objects of interest in the scene are identified and classified in the two-dimensional images.
14. These multiple two-dimensional images of the scene, and their accompanying object information, are then combined in a computer into a three-dimensional video database, or model, of the scene."

Col 9, Lines 52-61:

8. "Second, in advanced embodiments of the system, the computer is not limited to selecting from the three dimensional model a two-dimensional image that is, or that corresponds to, any of the images of the real-world scene as was imaged by any of the multiple video cameras. Instead, the computer may synthesize from the three-dimensional model a completely new two-dimensional image of the real-world scene as have been imaged by any of the multiple video cameras."

Col 10, Lines 8-13:

9. "In other words, in advanced embodiments of the invention the scene image shown may be a virtual image. Even if the image shown is a real image, the computer will still automatically select, and the display will still display, over time, those actual images of the scene as are imaged, over time, by different ones of the multiple video cameras."

Col 10, Lines 17-21:

10. "Forth, the user/viewer-specified criterion may be of a particular object in the scene. In this case the computer will combine the images from the multiple video cameras ... to generate a three-dimensional video model of the scene..."

Col 13, Lines 38-51:

11. "The 'viewer-selectable dynamic presentation' might be, for example, a viewer selected imaging of the quarterback. This image is dynamic in accordance that the quarterback should, by his movement during play, cause that, in the simplest case, the images of several different video camera should be successively selected of, in the case of such full virtual video as is contemplated by the present invention, that the quarterback's image should be variously dynamically synthesized by digital computer means. The football game is, of course, a dynamic event wherein the quarterback moves. Finally, the real-world source, camera, images that are used to produce the MPI video are themselves dynamic in accordance that the cameramen at the football game attempt to follow play."

Col 13, Lines 56-63:

12. "The experience of MPI video in accordance with the present invention may usefully be compared, and contrasted, with virtual reality. The term "virtual reality" commonly has connotations of (i) unreality, (ii) sensory immersion, and/or (iii) self-directed interaction with a reality that is only fantasy, or 'virtual'. The effect of the MPI video of the present invention differs from 'virtual reality' in all these factors..."

Col 14, Lines 4-6:

13. "In the second place, MPI video is presented upon a common monitor, or television set, and does not induce the viewer to believe that he or she has entered a fantasy reality."

Col 14, Lines 54-59:

14. "It is likely a better idea to construct a comprehensive video image database from quality images obtained from only a few strategically positioned cameras, and to then permit universal construction of customized views from this database, all as is taught by the present invention."

Col 17, Lines 3-8:

15. "In such cases, the MPI video system of the present invention will, by use of the environment model and video synthesis techniques, synthesize a virtual camera, and video image, so as to view a scene, and episode, or an entire presentation from a view-specified perspective."

Col 19, Lines 12-24:

4. "The ultimate response of the MPI video system is to synthesize the exact synthetic image, and image sequence, the viewer desires and demands. Even here, no image can be formed where no source image data exists, such as a view from below the playing field (i.e., from in the ground). Even a synthetic view that is normally acceptable, such as 'from the nose of the football is the vector direction of the 'movement of same' cannot be produced when, and at such times as, the football becomes 'buried', and obscured from view, under a pile up after the ball carrier is tackled. 'Weird' views in synthesized MPI video can be exciting, but, in accordance with their 'weirdness', are not always reliably capable of being successfully synthesized."

Col 37, Lines 59-64:

4. "It is the teaching of the present invention how to so construct from multiple two-dimensional video images a three-dimensional database, and how to so manage the three-dimensional database for the production of two-dimensional video images that not necessarily those images from which the database was constructed."

Col 39, Lines 17-25:

5. "First, the combination of multiple images, even video images, to generate a new image is called "morphing", and is, circa 1995, well known. One simple reason that the rudimentary system of the present invention does proceed to perform this "well known" step is that it is slow when performed on the engineering workstation on which the rudimentary embodiment of the present invention has been fully operationally implemented."

Col 41, Lines 25-30:

6. "In one, rudimentary, embodiment of the present invention, a virtual video camera, and a virtual video image, of a scene were synthesized in a computer and in a computer system from multiple real video images of the scene that were obtained by multiple real video cameras. This synthesis of a virtual video image was computationally intensive."

Col 42, Line17-23:

7. "Ultimate interactive control where each "fan" can be his own sports director is possible, but demands that considerable image data (actually, three-dimensional image data) be delivered to the "fan" either non-real time in batch (e.g. on CD-ROM), or in real time (e.g. by fiber optics), and, also, that the "fan" should have a powerful computer (e.g. and engineering workstation, circa 1995.)"

In summary, I believe that it is clear from both the descriptions in the current patent, as well as the descriptions contained within the original continued application entitled "Method for Representing Real-Time Motion over the Internet," that I teach the collection of 3-D body-points based upon the visual detection of strategically placed markers that ultimately form a mathematical database of the activities as opposed to a video database. This mathematical database of 3-D body-point locations over time is ideal for creating a animation of the game, which by its nature is "virtual" vs. "real." Jain discusses creating "real" images of the game either as direct feeds from "real" source cameras or as synthetic creations (i.e. morphing) of these "real" source camera images into a "realistic" new view, thereby representing a "virtual" source camera.

Furthermore, even if Jain were trying to work with markers to collect a similar 3-D body-point model to be used for animation, his "volume tracking" approach would be identical to the current state-of-the-art in this area; e.g. Motion Analysis and Vicon. As taught in my patent, this "volume" approach has significant drawbacks that have motivated the new designs and teachings of my application.

In addition to these three major areas of differentiation between the teachings and claims of my patent versus Jain, I would like to briefly touch and two additional areas of differentiation. Specifically:

Jain's teachings are designed to track objects with the assumption that they are "tracking-surface-bound." The following references from Jain's patent are sited in explanation.

Col 22, Line 63 – Col 23, Line 7:

"Three dimensional Scene Analysis. The purpose of scene analysis is to extract three-dimensional information from video frames captured by cameras. This process is performed in the following two stages:"

- "First, 2-D information is extracted. From each video frame, feature points such as players and field marks are extracted and a list of feature points is generated."
- "Second, 3-D information is extracted. From the two-dimensional description of the video frame, three-dimensional information in the scene, such as player position and camera status, is then extracted."

Here, Jain is saying that his first step is to find "feature points" in each 2-D image. Note that these feature points are not the same as the marked "body-points" that I teach. Examples of his "features" are an *entire player*, or field markers such as down lines. I am trying to identify up to 14 "feature" / body points on each player, which requires a wholly different technique with much better control over camera resolution than Jain has built into his system. This is why my patent teaches a second set of movable volume cameras that can find and stay zoomed in on each player or group of players to maintain the highest resolutions for finding the markers and the lowest occurrences of occlusions (i.e. player's blocking the view of other player's.)

I do not attempt to find "feature points" that are "field marks" because my overhead grid is pre-calibrated as if it were a single camera view looking at the playing field and already mapped to every square inch. There is no need to register each camera to "field marks," after all, neither

the field nor the overhead cameras move. Jain needs to do this precisely because he needs a way to coordinate one 2-D camera view to another 2-D camera view that may be partially overlapping, but from a significantly different perspective. If he can find the same "field marker(s)" in both images, than he can use this to help join the two 2-D views.

Col 27, Lines 42-47:

"All dynamic objects in the environment, including the robot, can be easily and accurately detected by (i) integrating motion information from the different cameras covering these objects, and, importantly to the invention, (ii) constraining the environment by analyzing only such motion as is constrained to be to a small set of known surfaces."

This "constraint" of assuming that all objects move along known surfaces is not present in my patent and *its absence* (i.e. my approach) is critical to efficient and accurate X, Y object tracking. There are two major reasons for this limitation of Jain. First, each camera is oriented at a perspective to the playing surface. Second, each camera's data is initially considered on its own to locate each object. By this second statement I mean that Jain analyzes each 2D image without coordination with, or even a guarantee of any overlapping images from a separate perspective. Since it is optically impossible to determine a three dimensional location of unknown objects in an unknown setting using a single view, Jain is forced to add limitations. (In other words, when you are at a perspective to a free floating object, you will need at least two overlapping images, each taken from a known and different perspective – Jain does not do this because it would significantly increase the number of cameras required.) Jain's approach is to put limits on the types of objects he is looking for and the setting in which he expects to find them.

First, he looks for the player's feet (now a known object with a pre-knowable basic size) and then he assumes that these feet will be touching the tracking surface (now a known setting.) How does he expect to track a helmet, a hand, an elbow, or the football itself? My patent is designed to find all of these "objects" in order to build the accurate 3-D body-point model. Simply but, Jain is not trying to track the same information, a critical distinction in our two patents. By assuming that all tracked objects must be touching a known surface, Jain can use a single camera oriented at an angle to that surface to make a reasonable assumption about that object's X, Y location. In my approach, the camera is substantially overhead and therefore substantially perpendicular to the tracking surface. After calibration to the tracking surface and accounting for any lens distortion, the location of the object in the image immediately yields its X, Y location – a very efficient approach.

Col 29, Lines 40-44:

"Then, by assuming that the lowest image point of a dynamic object is on the ground, the approximate position of the object on the ground plane is readily found. Positional information obtained from all view is assimilated and stored in the 2D grid representing the viewing area."

This assumption is a key difference between Jain's and my approach. Since my first set of area tracking cameras are substantially perpendicular to the tracking surface, they can easily determine X, Y locations of objects, without the need for assumptions like Jain's.

Col 29, Lines 29-35:

"For each single view, the preferred system is able to compute the three dimensional position of each dynamic object detected by its motion segmentation component. To do so, the (i) a priori information

about the scene and (ii) the camera calibration parameters are coupled with (iii) the assumption that all dynamic objects move on the ground surface."

Again, Jain limits the movement of the objects being tracked to the ground surface. As Jain later discusses, he will start to combine these "guesses" about each object's location from each different camera into his "intersecting frustum" approach. Unfortunately, this second order of processing is highly computational and introduces significantly delays in processing, in addition to being more susceptible to occlusions. My approach of using the overhead area tracking cameras is an immediate method for finding accurate X, Y object locations that does not require a second round of processing and is much less susceptible to object occlusions from a perspective view.

Col 11, Line 66 – Col 12, Line 7:

"Fig. 13 is a graphical illustration showing the intersection formed by the rectangular viewing frustum of each camera scene onto the environment volume in the GM-PPS portion of the MPI video system of the present invention; the filled frustum representing possible areas where the object can be located in the 3-D model while, by use of multiple views, the intersection of the frustum from each camera will closely approximate the 3-D location and form of the object in the environment model."

Reading Jain's approach carefully, it can be seen that each of his 2-D cameras make an approximate guess as to where each moving object is in X, Y space with respect to the tracking surface. Again, this "guess" requires limiting objects to a player's feet and then assuming those feet are touching the tracking surface. He then compares the projected angles from each camera through each object – this is the "intersection of the frustums." This comparison is of course a form of triangulation which may then significantly alter the "guess" of each object's X, Y location.

Col 28, Lines 45-52:

"For the case where an object is observed by more than one camera, the three-dimensional voxel representation is particularly efficacious. Here a dynamic object recorded on an image plane projects into some set of voxels. Multiple views of an object will produce multiple projections, one for each camera. The intersection of all such projections provides an estimate of the 3-dimensional form of the dynamic object as illustrated in Fig. 13 for an object seen by four cameras."

This is one of Jain's explanations of this "second round" of object location calculations. Notice that he talks about and object "seen by four cameras." Of course, he will need at least two views. All of this hints at the bigger problems and the reasons for my different approach. Once you try to have four cameras viewing the entire area, you immediately spread out the cameras resolution to such a point that it becomes harder to do the image analysis to find objects, and worse yet markers on parts of objects; i.e. there are just not enough pixels to go around. This means you really need to add cameras or you will not have sufficient resolution to find smaller objects accurately (like body locations, or an ice hockey puck). But as you add cameras, where should you put them? It is a significant amount of work for a computer to take all of the guesses from all of the cameras and sort them out into those likely to be the same object, and then to apply triangulation to see if they are actually aligned in all image planes.

Col 38, Line 56 – Col 39, Line 2:

"Other features extracted from the video images – such as football players and/or a football in motion – are much harder to extract, especially at high speeds and most especially in real time. To extract these moving features enters the realm of machine vision. Nonetheless that this portion of the system of the present invention is challenging, many simple machine solutions – ranging from fluorescently

bar-coded objects in the scene (e.g. players and football) to full-blown, state of the art machine vision programs – are possible and are discussed within this specification. In fact, with non-real-time video it is even possible – and quite practical – to have a trained human, or a squad of such, track each player or other object of concern through each video scene (e.g. a football play)."

I would like to respectfully say that Jain is "confessing" at this point that his "already working" system does not in fact automatically track objects. As later references will show, he absolutely requires human input to first locate and identify objects. And from here, the system attempts to intelligently follow each object. This distinction is not immediately obvious when reading his specification. Jain "glosses" over the difficulties of making any such system automatic and I submit grossly overstates that the necessary machine vision programs are "discussed within the specification."

Col 23, Lines 11-15:

"In the extraction of two-dimensional information, feature points are extracted from each video frame. Feature points include two separate items in the images. First, the players are defined by using their feet as feature points. Second, the field marks of the football field are used as feature points."

This reference reiterates Jain's approach of limiting the information on a tracked player to really be the location of his feet. There is no other detailed disclosure of how any other body points would be tracked. Clearly, a player's feet alone cannot be used as a means of identification. In my estimation, Jain's statements that his MPI system tracks and identifies objects are extremely misleading. As subsequent references will show, an operator first identifies and locates each player.

Col 23, Lines 23-58:

- "In the rudimentary, prototype, MPI system, the feature points are extracted by human-machine interaction. This process is currently carried out as follows. First, the system displays a video frame on the screen of Display 18 (shown in Fig. 1). A viewer, or operator, 14 locates some feature points on the screen and inputs required information for each feature point. The system reads image coordinates of the featured points and generates two-dimensional description."
- "This process results in two-dimensional description of a video frame that consists of a list describing the players and a list describing the field marks. The player descriptions include each player's name and the coordinates of each player's image. The field mark descriptions include the positions (in the three-dimensional world), and the image coordinates, of all the field, marks."
- "In the rudimentary embodiment of the MPI video system, all feature points are specified interactively with the aid of human intelligence. Many features can be detected automatically using machine vision techniques. See R. M. Haralick and L. G. Shapiro, op cit. The process of automatically detected features in arbitrary images is not trivial, however. It is anticipated, however, that two trends will help the process of feature point identification in MPI video. First, new techniques have recently been developed, and will likely continue to be developed, that should be useful in permitting the MPI video system to extract feature point information automatically. Future new techniques may include some bar-code like mechanism fore each player, fluorescent coloring on the players' helmets, or even some simple active devices that will automatically provide the location of each player to the system. It is also anticipated that many current techniques for dynamic vision and related areas may suitably be adapted fro the MPI video application."
- "Because the goal of the rudimentary, prototype, system is primarily to demonstrate MPI video, no extensive effort has been made to extract the feature points automatically."

Here, Jain is speculating on the likely technologies that will provide the basis for solutions that his patent has not taught. The difference of my two-camera-set approach vs. Jain's one fixed

volume arrangement, strongly support the use of machine vision to find multiple smaller features such as body points. Furthermore, my teachings do not require advanced machine vision techniques as Jain speculates, especially because the data collected is in a native orientation to the tracking area (set one of fixed cameras) and each object (set two of moveable cameras) that is conducive to simpler techniques.

Col 24, Line 58 – Col 25, Line 3:

"Ideally the scene analysis process just described should be applied to every video frame in order to get the most precise information about (i) the location of players and (ii) the events in the scene. However, it would require significant human and computational effort to do so in the rudimentary, prototype, MPI video system because feature points are located manually, and not by automation. Therefore, one key frame has been manually selected for every thirty frames, and scene analysis has been applied to the selected key frames. For frames in between, player position and camera status is estimated by interpolation between key frames by proceeding under the assumption that coordinate values change linearly between consecutive two key frames."

Again, this reality Jain is facing is a result of the raw video dataset as he has specified it to be collected, and not simply a result of insufficient processing power circa. 1995 – nor will it be "fixed" simply with better machine vision algorithms to come.

Col 25, Lines14-15:

"The system could prospectively be made more precise by considering the orientation of the player also."

My teachings show how to determine orientation.

Col 40, Lines 29-32:

"In the prototype MPI video system, much information was inserted manually by an operator. However to make MPI video practical for commercial use, this process should be automated as much as possible."

Clearly Jain understands that to be of "practical commercial use" systems such as he and I are proposing should be "automated as much as possible." Jain is conceding that he has not accomplished this nor taught specific solutions for making an automatic system. I respectfully submit that my teachings provide for an automatic system capable of real-time function with current technology and is the direct result of design choices as previously elaborated and as claimed.

And finally, the second additional area of differentiation between the teachings and claims of my patent versus Jain are a subtle but significant result of the type of data formats we are capturing and storing. Specifically:

Jain's stated goal is to allow the viewer to feel as if they can watch the game from virtually any angle. While this is also a goal of my system, Jain further wants this view to be identical in consistency to any normal, or "real," camera view. In fact, Jain spends considerable time discussing how his system can automatically choose between any of the actual video captured by a "real" camera or the synthesis of a "real" view for a "virtual camera" based upon the combination of 3-D data collected from the original "real" cameras. Obviously, as he also states, the viewer should not be able to tell the difference as the system automatically switches from "real" video to "synthesized" video. As the following excerpts from

Jain's patent will show, this goal is extremely ambitious and has severe drawbacks. While my patent does teach how to collect "real" video for a viewer, I do not teach a method of using this video to synthesize new views from "virtual cameras." Furthermore, my current set of claims is not directed towards this "real" video but rather other aspects of my teachings which show how to collect a 3-D body point model that can then be used to animate any desired view. Obviously, this animation is not real-video.

The following references from Jain's patent are sited in explanation.

Col 2, Lines 7-17:

> "In an extension of the present invention the image presented to the viewer will be seen to be a virtual image that is not mandated to correspond to any real camera nor to any real camera image. A viewer may thus view a video or television of a real-world scene from a vantage point (i.e., a perspective on the video scene), and/or dynamically in response to objects moving in the scene and/or events transpiring in the scene, in manner that is not possible in reality. The viewer may, for example, view the scene from a point in the air above the scene, or from the vantage point of an object in the scene, where no real camera exists or even, in some cases, can exist."

As previously discussed, Jain's goal is to take the video captured by a few video cameras and to be able to combine this to form nearly any desired view, as if there were a real camera in existence at that view. There are some systems known to the inventor that also perform this type of function. These systems specify a tight alignment of "real" source cameras, one at every six degrees around a circle. This type of arrangement ensures that enough raw video is collected to ensure realistic new "virtual" views. Still, even these systems have severe limitations because of this circular arrangement requirement.

Col 13, Lines 38-51:

> "The 'viewer-selectable dynamic presentation' might be, for example, a viewer selected imaging of the quarterback. This image is dynamic in accordance that the quarterback should, by his movement during play, cause that, in the simplest case, the images of several different video camera should be successively selected of, in the case of such full virtual video as is contemplated by the present invention, that the quarterback's image should be variously dynamically synthesized by digital computer means. The football game is, of course, a dynamic event wherein the quarterback moves. Finally, the real-world source, camera, images that are used to produce the MPI video are themselves dynamic in accordance that the cameramen at the football game attempt to follow play."

Here, Jain is saying that his system will choose either existing views to follow the quarterback, or will synthesize a new view from those existing views. As will be seen, there are significant processing and data storage requirements necessary to even attempt this functionality with four cameras, let alone the thirty or more cameras that would more than likely be some lowest acceptable number of original source views.

Col 14, Lines 44-54:

> "It has been hypothesized that the Internet, in particular, may expand in the future to as likely connect smart machines to human users, and to each other, as to communicatively interconnect more and more humans, only. Customized remote viewing can certainly be obtained by assigning every one his or her own remotely-controllable TV camera, and robotic rover. However, this scheme soon breaks down. How can hundreds and thousands of individually-remotely-controlled cameras jockey for position and for viewer-desired vantage points at a singe event...?"

Jain is arguing that it is simply unfeasible to service the "hundreds and thousands of individually-remotely-controlled cameras" that might be necessary to let everyone be their own "director" choosing their own camera view.

Col 14, Lines 54-59:

"It is likely a better idea to construct a comprehensive video image database from quality images obtained from only a few strategically positioned cameras, and to then permit universal construction of customized views from this database, all as is taught by the present invention."

Jain's solution is to replace the physical camera bottleneck with an equally challenging approach of servicing thousands and thousands of individual viewer requests for uniquely synthesized views all from a central server. As Jain himself further admits, this is no "trivial" task.

This approach is significantly different from my patent where I teach collecting a mathematical model, rather than video data, which can easily be transmitted to each viewer even with today's bandwidths. The viewer's own system then recreates the desired view using their local computer or TV set-top box, again using the equivalent of today's video gaming hardware.

Col 7, Lines 7-16:

"The present invention will be seen to manage a very great amount of video data. A three-dimensional video model, or database is constructed. For any sizable duration of video (and a sizable length thereof may perhaps not have to be retained at all, or at least retained long), this database is huge. More problematical, it takes very considerable computer "horse-power" to construct this database – however long its video data should be held and used."

However, with Jain's "likely better idea" there are significant data storage and processing problems as he admits.

Col 39, Lines 17-25:

"First, the combination of multiple images, even video images, to generate a new image is called "morphing", and is, circa 1995, well known. One simple reason that the rudimentary system of the present invention does proceed to perform this "well known" step is that it is slow when performed on the engineering workstation on which the rudimentary embodiment of the present invention has been fully operationally implemented."

Again, Jain found via experience that synthesizing video images was significantly costly in terms of processing power. I would further submit that the quality of these synthesized images would be significantly degraded over their "real" sources and that as the industry further moves into HDTV, the possibility of using Jain's approach for generating realistic synthesized video from any perspective even for a single view would be daunting – even without the real-time restriction. My solution avoids these issues by instead relying upon advances in animation that continue to deliver more and more realistic views, including moving actual faces of players.

Col 42, Line17-23:

"Ultimate interactive control where each "fan" can be his own sports director is possible, but demands that considerable image data (actually, three-dimensional image data) be delivered to the "fan" either non-real time in batch (e.g. on CD-ROM), or in real time (e.g. by fiber optics), and, also, that the "fan" should have a powerful computer (e.g. and engineering workstation, circa 1995.)"

Seeing this formidable problem of a centralized server creating all of the requested synthesized views, Jain proposes sending the collected and assembled 3-D video database to each fan's remote system thereby letting them create their own camera views with their own computer hardware. This is simply not feasible. The data requirements for his 3-D database, using just four cameras, could not be realistically transmitted over any medium including fiber optic in anything close to real-time. Let alone being transmitted to thousands and thousands of viewers – just so they could create their own view. My patent understands these real problems and is therefore directed towards animation as the means to provide a viewer with unlimited viewing angles.

Col 41, Lines 25-30:

"In one, rudimentary, embodiment of the present invention, a virtual video camera, and a virtual video image, of a scene were synthesized in a computer and in a computer system from multiple real video images of the scene that were obtained by multiple real video cameras. *This synthesis of a virtual video image was computationally intensive*." (Italics added)

Another excerpt from Jain where he is acknowledging the significant drawbacks of his proposed technology.

Col 41, Lines 14-23:

"As access to data from more and more cameras is permitted, the storage requirements for MPI video will increase significantly. Where and how to store this video data, and how to organize it for timely retrieval, is likely to be a major issue for expansion and extension of the MPI video system. In the prototype system, the single most critical problem has been the storage of data. Future MPI video will continue to put tremendous demands on the capacity and efficiency of organization of the storage and database system."

Ultimately, I believe that Jain's approach is significantly hampered by the storage and processing requirements of creating, storing and synthesizing from his 3D video database. Again, my solution is significantly different and avoids all of these limitations.

In summary, I submit that my teachings are significantly different from Jain's for at least the following reasons:

1. I teach a first set of fixed area tracking cameras whose data streams controllably affects a second set of movable volume tracking cameras. Jain teaches a single set of fixed volume tracking cameras.
2. I teach that the first set of fixed area tracking cameras should be aligned in a regular matrix where each camera is overhead of, and substantially perpendicular to, the tracking area as well as partially overlapping with their neighbors to form a single contiguous co-planer view. Jain alternatively specifies that the cameras are not to be in a regular arrangement, they are not to be substantially overhead, the do not necessarily overlap with their neighbors and they are not perpendicular to the tracking area.
3. I teach the construction of a 3D body-point model that is mathematical in nature and used as the basis for animation. Jain teaches the creation of a 3D video database that is the storage of realistic video supporting synthetic morphing of real images.
4. I teach the accurate X, Y tracking of objects without the prior assumption that they are "surface bound." Jain's first guess location assumes each object is surface bound. Upon examining each 2D image from each area tracking camera, I have sufficient information to accurately determine each object's X, Y location and orientation. Jain does not determine orientation, and requires a "second

pass" of comparing the frustums calculated during the first pass in order to accurately determine the X, Y locations.

5. I teach pushing the smaller set of 3D body-point data to the remote viewer for animation on their computer or equivalent. Jain teaches the creation of a very large 3D video dataset that is not easily transmitted to the viewer in real-time and therefore must be processed locally at the server, creating significant processing requirements beyond any reasonably foreseeable future computing technology.

I respectfully request that you allow my revised claims as clearly patentable over the cited reference of Jain alone, and of Jain in combination with the secondary reference of Leis.

I thank you for your consideration in these matters.

Sincerely,

James A. Aman

This communication was e-mailed to Senfi, Behrooz [Behrooz.Senfi@USPTO.GOV] on 10/17/05. It was also mailed Post Office To Addressee from Harleysville, Pennsylvania on 10/17/05 using label EQ006550 92US.